

#3

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
OHTA et al.)
Application Number: To Be Assigned)
Filed: Concurrently Herewith)
For: QUERY MODIFICATION SYSTEM FOR INFORMATION)
RETRIEVAL)

10/076400
02/19/02

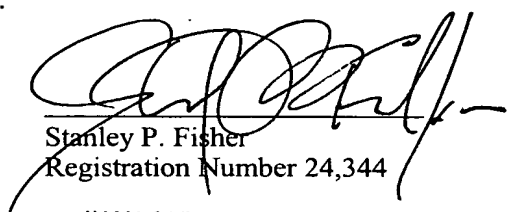
Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**NOTICE OF PRIORITY
UNDER 35 U.S.C. 119
AND THE INTERNATIONAL CONVENTION**

Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of June 29, 2001, the filing date of the corresponding Japanese patent priority application 2001-198757.

A certified copy of corresponding Japanese patent application 2001-198757 is being submitted herewith. The Examiner is most respectfully requested to acknowledge receipt of the certified copy in due course.


Stanley P. Fisher
Registration Number 24,344

JUAN CARLOS A. MARQUEZ
Registration No. 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200

February 19, 2002

(Translation)

PATENT OFFICE
JAPANESE GOVERNMENT



This is to certify that the annexed is a true copy of
the following application as filed with this Office.

Date of Application: June 29, 2001

Application Number: Japanese Patent Application
No. 2001-198757

Applicant(s): Hitachi, Ltd.

November 9, 2001

Commissioner,
Patent Office

Kozo Oikawa (seal)

Certificate No. 2001-3099405

日 本 国 特 許 庁
JAPAN PATENT OFFICE

JP01 U.S. PTO
10/076400
02/19/03

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日
Date of Application:

2001年 6月29日

出 願 番 号
Application Number:

特願2001-198757

出 願 人
Applicant(s):

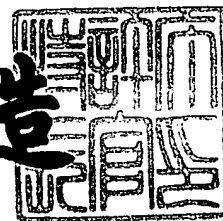
株式会社日立製作所

CERTIFIED COPY OF
PRIORITY DOCUMENT

2001年11月 9日

特 許 庁 長 官
Commissioner,
Japan Patent Office

及 川 耕 造



出証番号 出証特2001-3099405

【書類名】 特許願

【整理番号】 H100449

【提出日】 平成13年 6月29日

【あて先】 特許庁長官 殿

【国際特許分類】 G06F 17/30

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社
日立製作所 中央研究所内

【氏名】 大田 佳宏

【発明者】

【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社
日立製作所 中央研究所内

【氏名】 西川 哲夫

【発明者】

【住所又は居所】 東京都千代田区神田駿河台四丁目 6 番地 株式会社 日
立製作所 ライフサイエンス推進事業部内

【氏名】 井原 茂男

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社 日立製作所

【代理人】

【識別番号】 100091096

【弁理士】

【氏名又は名称】 平木 祐輔

【手数料の表示】

【予納台帳番号】 015244

【納付金額】 21,000円

【その他】 国等の委託研究の成果に係る特許出願（平成 1 1 年度新
エネルギー・産業技術総合開発機構（再）委託研究、産

業活力再生特別措置法第 3 0 条の適用を受けるもの)

【提出物件の目録】

【物件名】	明細書	1
【物件名】	図面	1
【物件名】	要約書	1
【プルーフの要否】	要	

【書類名】 明細書

【発明の名称】 情報検索システム及びサーバ

【特許請求の範囲】

【請求項 1】 データベースから情報を検索するための情報検索システムにおいて、

問い合わせ用の情報を入力するための入力画面を表示する手段と、

入力された問い合わせ用の情報から構築した問い合わせ概念を複数のキーワードと各キーワードの重みとを含むクエリーベクトルとして表示するクエリーベクトル表示手段とを備えることを特徴とする情報検索システム。

【請求項 2】 請求項 1 記載の情報検索システムにおいて、

前記入力画面は、情報をテキスト形式で保存しているファイル名、自然言語による文や句、公共データベースの ID 番号、URL、既に登録済みの問い合わせ概念の識別情報のいずれか又はその組み合わせによって問い合わせ用の情報を入力することができ、

前記クエリーベクトル表示手段は、前記入力画面に入力された問い合わせ情報を統合して生成したクエリーベクトルを表示することを特徴とする情報検索システム。

【請求項 3】 請求項 1 記載の情報検索システムにおいて、前記クエリーベクトル表示手段に表示されたクエリーベクトルを編集する手段を備えることを特徴とする情報検索システム。

【請求項 4】 請求項 3 記載の情報検索システムにおいて、前記クエリーベクトルを編集する手段は、前記クエリーベクトル表示手段に表示されたキーワードを、指定した重み以上のキーワードだけに制限する手段、あるいは、指定した順位までの重みの大きなキーワードだけに制限する手段を有することを特徴とする情報検索システム。

【請求項 5】 請求項 3 記載の情報検索システムにおいて、前記クエリーベクトルを編集する手段は、前記クエリーベクトル表示手段に表示されたキーワードの重みを個別に変更する手段を有することを特徴とする情報検索システム。

【請求項 6】 請求項 1 記載の情報検索システムにおいて、検索結果として

、一方の軸に検索された文献をスコアの高い順に配置し、他方の軸にクエリーベクトルの要素である複数のキーワードを配置し、各文献とキーワードとの交点に各文献における前記キーワードのスコアを配置した表を表示する手段を備えることを特徴とする情報検索システム。

【請求項 7】 請求項 1 記載の情報検索システムにおいて、検索結果として得られた文献中で前記クエリーベクトル中のキーワードと共起する単語を抽出し一覧表示する手段と、当該一覧表示された単語の中で指定された単語を前記問い合わせ用の情報に追加する手段とを備えることを特徴とする情報検索システム。

【請求項 8】 請求項 1 記載の情報検索システムにおいて、検索された文献をスコア順位の高い順に一覧表示する検索結果表示手段と、前記検索結果表示手段に表示された文献の中で指定された文献を前記問い合わせ用の情報に追加する手段を備えることを特徴とする情報検索システム。

【請求項 9】 請求項 7 又は 8 記載の情報検索システムにおいて、変更された問い合わせ用の情報に基づいて問い合わせ概念を再構築し、複数のキーワードと各キーワードの重みとを含むクエリーベクトルとして表示する手段を備えることを特徴とする情報検索システム。

【請求項 10】 クライアントから送信されてきた問い合わせ用の情報から複数のキーワードと各キーワードの重みとを含むクエリーベクトルを生成する手段と、

前記クエリーベクトルを表示した画面をクライアントに送信する手段と、

情報検索のために前記クエリーベクトルをデータベースに送信する手段と、

前記データベースによる検索結果を表示した画面をクライアントに送信する手段とを含むことを特徴とするサーバ。

【請求項 11】 請求項 10 記載のサーバにおいて、検索結果として得られた文献中で前記クエリーベクトル中のキーワードと共起する単語を抽出する手段と、抽出した単語の一覧表示画面をクライアントに送信する手段と、前記一覧表示画面の中でクライアントが指定した単語を前記問い合わせ用の情報に追加してクエリーベクトルを再構成する手段とを備えることを特徴とするサーバ。

【請求項12】 請求項10記載のサーバにおいて、前記データベースによって検索された文献をスコア順位の高い順に一覧表示した検索結果表示画面をクライアントに送信する手段と、前記検索結果表示画面に表示された文献の中でクライアントが指定した文献を前記問い合わせ用の情報に追加してクエリーベクトルを再構成する手段とを備えることを特徴とするサーバ。

【請求項13】 請求項1～9のいずれか1項記載の情報検索システムをコンピュータに実現させるためのプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明はインターネット上の情報検索に係わり、例えば生命科学分野の文献を検索し、それに付随した情報を表示する情報検索システム及びサーバに関する方法に関する。

【0002】

【従来の技術】

情報検索の研究には半世紀近い歴史があるが、その根幹には学術情報をどのように配布するか、あるいは収集するかという問題意識があった。したがって、情報検索の検索対象は、書籍や学術論文などのように均質で閉じた世界のものが中心であった。これに対して、1990年代に爆発的な普及をとげたインターネットは情報検索の研究分野に大きなインパクトを与えた。インターネット上の情報は、変化の速度、絶対量、非永続性、非均質性、媒体の多様性、開放性などの点で従来の情報検索の研究が対象としていた情報とは異質である。このように質的に異なる検索対象を扱うためには、これまでの情報検索で用いられてきた手法では必ずしも十分ではない。最近、情報検索の研究分野が活性化しているのもインターネットの普及によるところが多い。

【0003】

より知的で性能の良い情報検索システムが求められているインターネット上の検索サービスは、大きくYahoo! (<http://www.yahoo.com/>) のようなディレクトリ型と、Alta Vista (<http://www.altavista.com>) やGoogle (<http://www.google>)

e.com/) のようなロボット型に分類できる。ディレクトリ型検索サービスでは、URLを人手により分野別に分類する方式を取っており、データ量が少ない反面、人手で索引や要約を作成するため、索引と要約の信頼性が高いといった特徴を持つ。一方、ロボット型検索サービスでは、WWWロボットやスパイダーと呼ばれるWeb探索プログラムを用いて、インターネット上で見つけることの出来るWWWサーバ上の情報を定期的に収集し、その情報の索引付けを行っており、情報量が多いという利点を持つ。ロボット型検索サービスのGoogleでは、従来のテキストに対する索引付けを行い、類似度を計算することで行ってきた情報検索の手法だけでなく、そのページに関するリンク情報をもとに算出したPage Rankという要素を加味することで、情報検索システムとしての性能を向上させている。

【0004】

このような従来の手法だけではなく、様々な試みを取り入れる動きは多く、特に、インターネット上のリソースでも、分野を限定している場合のみ適用可能な手法なども開発されている。生命科学分野の情報発信のサイトである米国National Center for Biotechnology Information (NCBI) の文献データベースであるPubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>) に対してもそのような試みがなされている。そこでは、問い合わせにおいて与えられた遺伝子名をもとに、その遺伝子に関して最もよく説明されている文献を抽出し、その文献との類似度の高い文献を検索できるという試みである。生命科学の分野においては、ヒトゲノムプロジェクトの進展(2000年7月にドラフトシーケンス完了)に伴い、その関連論文が日々増大しているのが現状である。PubMedにおいても、日々複数の論文が新規登録され、更新されている。このような状態の検索対象から、ユーザごとの要求に適した形で情報を抽出する作業は、いまだ困難な状態であると言える。

【0005】

ここで、情報検索とは、ユーザの与えるクエリに適合する文書を文書集合の中から見つけ出すことである。クエリとは、ユーザが問題を解決するために必要と感じている情報への要求を具体化したものであり、直接、情報検索システムに入力することのできる形式のものである。情報検索システムとは、ユーザからのク

エリを受け、計算機がクエリに適合する文書を文書集合の中から見つけ出し、ユーザに提示するという一連のシステムである。計算機における情報検索システムでは、検索対象となる文書集合とユーザから与えられたクエリは、計算機の内部で扱えるようにするために、計算機の内部表現へと変換される。その上で、両者を比較することで、計算機は検索を行うことになる。検索対象となる文書集合やユーザから入力されたクエリを計算機上で扱える内部表現に変換するための処理を、索引付けと呼ぶ。文書は文章の集まりであり、文章は単語の集まりであるというのが、索引付けの基本的な考えであり、このときの最小単位となる単語などを索引語と呼ぶ。この考えに基づき、各文書 d_i はそれを構成する各索引語 t_j の出現頻度 w_{ij} をもって、式(1.1)のようなベクトルとして表現することができる。

【0006】

【数1】

$$d_i = \begin{pmatrix} w_{i1} \\ w_{i2} \\ \vdots \\ w_{i2} \\ \vdots \\ w_{iM} \end{pmatrix} \quad \text{式(1.1)}$$

【0007】

索引付けの処理においては、一般に次のような処理を行う。

- (1) 不要語リストを参照して文書中の不要語を削除
- (2) 接辞処理
- (3) 語の頻度をもとにして索引語に重み付け

【0008】

索引付けの主な役割は、文書の中からその文書を特徴付ける索引語を漏れなく抽出することであるが、さらに抽出した索引語がその文書にどれだけ密接に関係しているかを索引語の重要度として索引語に付与することもできる。抽出した索引語にその索引語の重要度を表す尺度を与えることを索引語の重み付けと呼ぶ。

索引語の重み付けの最も簡単なものは、その索引語が文書の中で何回使われたかという頻度そのものを用いる場合である。ある文書 d_i を構成する各索引語 t_j の出現頻度を w_{ij} とすると、各文書としては式(1.1)のようなベクトルとして見ることができるが、ここでは、式(1.2)のような行列を考える。つまり、各行はその索引語の文書にわたる分布を表し、各列はその文書内の索引語の分布を表している。

【0 0 0 9】

【数2】

$$A = \begin{matrix} & d_1 & d_2 & \cdots & d_M \\ \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{matrix} & \begin{bmatrix} w_{11} & w_{21} & \cdots & w_{M1} \\ w_{12} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ w_{1N} & \cdots & \cdots & w_{MN} \end{bmatrix} \end{matrix} \quad \text{式(1.2)}$$

【0 0 1 0】

このように検索対象となる文書集合を行列として計算機の内部に持つことは、後のクエリとの比較、つまり実際の検索において効率が良い。

上記までは、検索対象となる文書の内部表現について説明した。次に、ユーザから入力されたクエリの内部表現について説明する。クエリの入力は、索引語の直接入力を扱う。この索引語の集合を上記の検索対象と同様に、計算機の内部表現へと変換することになる。クエリについても、基本的には上記までの検索対象と同様の処理を行う。つまり、不要語の処理、接辞処理、重み付けを行うのである。ただし、クエリは、文書集合のように複数あるわけではなく、1回の検索に対しては1つのクエリのみということになるので、式(1.2)のような行列としてではなく、次の式(1.3)のように、クエリ q は各索引語 t_j の出現頻度 w_{qj} を要素として持つベクトルとして与えられることとなる。

【0 0 1 1】

【数 3】

$$q = \begin{pmatrix} w_{q1} \\ w_{q2} \\ \vdots \\ w_{qj} \\ \vdots \\ w_{qM} \end{pmatrix} \quad \text{式(1.3)}$$

【0012】

ここまでで、検索対象となる文書集合とユーザから入力されたクエリは、それぞれ索引語とその頻度によって同様の形式の内部表現へと変換された。それを用いた文書とクエリの比較によって検索を行うのであるが、その比較方法である検索モデルはこれまでに数多く提案されている。その代表的な例には、ブーリアンモデル、ベクトル空間モデル、確率モデル、ファジィ集合モデル、拡張ブーリアンモデル、ネットワークモデル、クラスタモデル等がある。

【0013】

文書とクエリとを比較する検索モデルの最も簡単なものは、ブーリアンモデルである。ブーリアンモデルでは、クエリで用いられた索引語と完全一致する索引語を含む文書を抽出するだけというもので、論理演算によって簡単に求まる。また、処理の高速化の技術も考案されており、実用向きである。ただし、この手法では検索結果に順位をつけることができないため、一般には他の方法と併用されることが多い（徳永健伸：“情報検索と言語処理，言語と計算5”，東京大学出版会，1999）。

【0014】

今回とりあげる検索システムのベースとなる手法のベクトル空間モデルでは、各文書を式(1.2)の各列を取り出した列ベクトルとし、それと同次元である式(1.3)のクエリベクトルとの類似度を測る。この類似度により、検索結果に順位をつけることができるのである。ベクトル同士の類似度は、その余弦(式(1.4))によって計算されることが多い。これは、余弦を用いることで、検索の性能が上がるという実験的な報告を受けてのものである。余弦を用いることは、両ベクトルの

張る角度を見ることになり、また、ベクトルのノルムは無視されることになるので、値が1に近いほど、その類似度が高いということになる。ただし、ベクトル空間モデルは、全ての文書との類似度計算をするため、一般にはブーリアンモデルにより検索対象を絞り込んでから使うことが多い。

【0015】

【数4】

$$\delta(d_i, d_j) = \frac{\sum_{k=1}^M w_{ik} \cdot w_{jk}}{\sqrt{\sum_{k=1}^M w_{ik}^2 \times \sum_{k=1}^M w_{jk}^2}} \quad \text{式(1.4)}$$

【0016】

【発明が解決しようとする課題】

本発明は、例えばPubMedのような生命科学分野の文献データベースを活用し、ユーザの要求する情報をよりの確に、より分かりやすく提供するための情報検索システムを提供することを目的とする。

【0017】

【課題を解決するための手段】

本発明では、ユーザの要求をより高度に実現するために、問い合わせの生成、検索結果の表示、検索結果の問い合わせへのフィードバックなどにおいて、問い合わせ用の情報を入力するための画面を表示する手段と、入力された問い合わせ用の情報から構築した問い合わせ概念をクエリーベクトルとして表示する手段、及び、問い合わせ概念の編集を可能とする手段の実装を行った。具体的には以下の機能があげられる。

【0018】

- (1) 問い合わせは、様々な形態のものを採用できるようにすること。
- (2) 検索途中の経過を表示しつつ、それに対してもアクションできるようにすること。
- (3) 検索結果の詳細から、様々な情報を引き出せるようにすること。
- (4) 検索結果から、問い合わせへの様々なフィードバックを行えるようにすること。

【0019】

本発明による情報検索システムあるいはサーバは、以下の特徴を有する。

(1) データベースから情報を検索するための情報検索システムにおいて、問い合わせ用の情報を入力するための入力画面を表示する手段と、入力された問い合わせ用の情報から構築した問い合わせ概念を複数のキーワードと各キーワードの重みとを含むクエリーベクトルとして表示するクエリーベクトル表示手段とを備えることを特徴とする情報検索システム。

【0020】

(2) (1) 記載の情報検索システムにおいて、前記入力画面は、情報をテキスト形式で保存しているファイル名、自然言語による文や句、公共データベースPubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>) のID番号、URL、既に登録済みの問い合わせの識別情報のいずれか又はその組み合わせによって問い合わせ用の情報を入力することができ、前記クエリーベクトル表示手段は、前記入力画面に入力された問い合わせ情報を統合して生成したクエリーベクトルを表示することを特徴とする情報検索システム。

公共データベースのID番号としては、例えば公共データベースPubMed (<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed>) のUI番号がある。

【0021】

(3) (1) 記載の情報検索システムにおいて、前記クエリーベクトル表示手段に表示されたクエリーベクトルを編集する手段を備えることを特徴とする情報検索システム。

(4) (3) 記載の情報検索システムにおいて、前記クエリーベクトルを編集する手段は、前記クエリーベクトル表示手段に表示されたキーワードを、指定した重み以上のキーワードだけに制限する手段、あるいは、指定した順位までの重みの大きなキーワードだけに制限する手段を有することを特徴とする情報検索システム。

【0022】

(5) (3) 記載の情報検索システムにおいて、前記クエリーベクトルを編集する手段は、前記クエリーベクトル表示手段に表示されたキーワードの重みを個別

に変更する手段を有することを特徴とする情報検索システム。

(6) (1) 記載の情報検索システムにおいて、検索結果として、一方の軸に検索された文献をスコアの高い順に配置し、他方の軸にクエリーベクトルの要素である複数のキーワードを配置し、各文献とキーワードとの交点に各文献における前記キーワードのスコアを配置した表を表示する手段を備えることを特徴とする情報検索システム。

【 0 0 2 3 】

(7) (1) 記載の情報検索システムにおいて、検索結果として得られた文献中で前記クエリーベクトル中のキーワードと共起する単語を抽出し一覧表示する手段と、当該一覧表示された単語の中で指定された単語を前記問い合わせ用の情報に追加する手段とを備えることを特徴とする情報検索システム。

(8) (1) 記載の情報検索システムにおいて、検索された文献をスコア順位の高い順に一覧表示する検索結果表示手段と、前記検索結果表示手段に表示された文献の中で指定された文献を前記問い合わせ用の情報に追加する手段を備えることを特徴とする情報検索システム。

【 0 0 2 4 】

(9) (7) 又は (8) 記載の情報検索システムにおいて、変更された問い合わせ用の情報に基づいて問い合わせ概念を再構築し、複数のキーワードと各キーワードの重みとを含むクエリーベクトルとして表示する手段を備えることを特徴とする情報検索システム。

(10) クライアントから送信されてきた問い合わせ用の情報から複数のキーワードと各キーワードの重みとを含むクエリーベクトルを生成する手段と、前記クエリーベクトルを表示した画面をクライアントに送信する手段と、情報検索のために前記クエリーベクトルをデータベースに送信する手段と、前記データベースによる検索結果を表示した画面をクライアントに送信する手段とを含むことを特徴とするサーバ。

【 0 0 2 5 】

(11) (10) 記載のサーバにおいて、検索結果として得られた文献中で前記クエリーベクトル中のキーワードと共起する単語を抽出する手段と、抽出した単

語の一覧表示画面をクライアントに送信する手段と、前記一覧表示画面の中でクライアントが指定した単語を前記問い合わせ用の情報に追加してクエリーベクトルを再構成する手段とを備えることを特徴とするサーバ。

(12) (10) 記載のサーバにおいて、前記データベースによって検索された文献をスコア順位の高い順に一覧表示した検索結果表示画面をクライアントに送信する手段と、前記検索結果表示画面に表示された文献の中でクライアントが指定した文献を前記問い合わせ用の情報に追加してクエリーベクトルを再構成する手段とを備えることを特徴とするサーバ。

(13) (1) ~ (9) のいずれか 1 項記載の情報検索システムをコンピュータに実現させるためのプログラム。

【0026】

【発明の実施の形態】

以下、図面を参照して本発明の実施の形態を説明する。

本発明の情報検索システムでは、クエリと文書中の索引語が一致することに基づいて検索を行う。したがって、本来、同一であるべき索引語が言語の多様性によって不一致になると、検索すべき文書が検索できなくなってしまう。言語表現の多様性には語形の多様性と語選択の多様性がある。語形の多様性の問題を解決するために接辞処理を行う。ここでは、もう一つの多様性、語選択の多様性を考える。語選択の多様性とは、ある概念を表現するのに様々な語を用いて表現できるということである。この語選択の多様性の問題を解決するためには、以下の2つの方法が考えられている。

(1) 同じ概念を表す表現は全て同一の記号に変換する。

(2) クエリ中に含まれる表現をそれと同じ概念を表す全ての表現の集合と置き換える。

【0027】

(1)の方法は、語形の多様性を扱うために接辞処理を行ったように、表層的には違うが本来同じものを全て同一の記号に縮退するというアプローチで、“road”、“street”、“way”などを“@ROAD”のような概念を表す記号に変換する方法である。(2)の方法は、ある一つの表現をそれと同じ概念を表す全ての表現に拡張する

アプローチで、クエリ中に、"road"とあれば、それを"road"、"street"、"way"というように置き換える方法である。(Bruce R. Schatz, Eric H. Johnson, Pauline A. Cochrane: "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval", Proceeding Digital Libraries '96: 1st ACM International Conference on Research and Development in Digital Libraries, March 20-23 1996 in Bethesda, MD.)

【0028】

ここではまず、図1を用いて問い合わせ概念の生成方法について説明する。画面101は問い合わせ概念の生成用の画面であり、ファイル名入力用フォーム102、自然言語入力用フォーム103、UI番号入力用フォーム104、URL入力用フォーム105、前回作成して保存しておいた問い合わせ概念の読み出し用フォーム106を持ち、問い合わせ概念の生成処理の実行用ボタン107を持つ。問い合わせ用の情報として、既にテキスト形式のファイルで用意されたものを入力する際は、ファイル名入力用フォーム102にそのファイルのファイル名をフルパスで入力する。同様に、問い合わせ用の情報として自然言語を入力する際は、自然言語入力用フォーム103に自然言語を記述し、Medline IDであるUI番号を入力する際は、UI番号入力用フォーム104にUI番号を記述し、インターネット上のあるページを入力とする際は、URL入力用フォーム105にURLを記述する。既に登録してある問い合わせを入力する際は、読み出し用フォーム106を用いて登録済みの問い合わせの識別情報を記述する。

【0029】

一連の操作の後、問い合わせ概念の生成処理の実行用ボタン107を押すことで、指定されたものについての問い合わせ概念、及びそれらを統合した問い合わせ概念をクエリーベクトルとして生成する。ここで統合した問い合わせ概念は、各フォーム毎のクエリーベクトルの足し算で作成される。クエリーベクトルが生成されると、問い合わせ概念の詳細を表示する画面108が表示される。画面中、109はクエリーベクトルのキーワードのリストを表す。110はタグのリストを表す。ここでタグとは、キーワードの属する分類クラスを表している。例えば、キーワ

ード “glucocorticoid” はタンパク質名なので “PROTEIN” タグが割り当てられている。この画面108は、問い合わせ概念をリスト109のキーワード、リスト110のタグ、リスト111の重みをもって表現し、表示している。

【0030】

図2の画面201、及び、画面208は問い合わせ概念の表示例を表している。画面201では、重みが「0.1」以上のキーワードで、かつ、重みの値が上位10件以内のものだけを表示している。件数入力フォーム203を用いて、上位何件までを表示するかを記述し、重み入力フォーム204を用いて、重みがいくつ以上のキーワードを表示するかを記述する。件数入力フォーム203、及び、重み入力フォーム204を記述後、表示を更新するための表示ボタン202を押すことで、上記条件を満たす問い合わせ概念のキーワードのみが一覧として表示される。一覧は、前述の通りリスト205のキーワード、リスト206のタグ、リスト207の重み、以上3つの要素を表示する。画面208では、重みが「0.01」以上のキーワードで、かつ、重みの値が上位100件以内のものだけを表示している。このように、件数入力フォーム203、重み入力フォーム204、及び、表示ボタン202を用いることで、問い合わせ概念の詳細を確認することができる。

【0031】

次に、図3により問い合わせ概念の詳細確認について説明する。画面301は、問い合わせ概念の表示画面である。ここで、リスト302のキーワード、リスト303のタグ、リスト304の重みについては、前述の通りである。この画面301が表示されている状態で、リスト302のキーワードのうち、追加情報を知りたいキーワードをクリックするとサブウィンドウ310が開き、そのキーワードについての追加情報をあらかじめシステムに登録しておいたオンライン上のデータベースで検索することができる。

【0032】

画面305、及び、画面308は、キーワード “glucocorticoid” をクリックしたとき開いたサブウィンドウ310に表示されたデータベースで検索した結果を表示したものである。画面305は、タンパク質についてのデータベース(PDB)を検索した結果の画面で、リスト306に挙げられたものが検索結果である。3次元グラフィック

307は、選択したタンパク質の立体構造を表し、角度変更や拡大縮小を用いて細部を確認することができる。また、画面308は、配列データベース (Genbank) を検索した結果の画面で、リスト309は検索結果の名前と配列の詳細を記述したものである。

また、サブウィンドウ310に表示されている "modify" をクリックすると、weight変更画面が現れ、そこに数値を入力することで、そのサブウィンドウ310を開いたキーワードの重みの数値を変更することができる。

【0033】

次に、図4によりキーワードの追加について説明する。画面401は、前述の問い合わせ作成画面である。この画面401の "Suggetion" ボタン407をマウスでクリックすることにより展開された画面402は、文献を解析することによって予測した問い合わせ概念に追加すべきキーワードの候補となるものの一覧を、ユーザに提示する表示画面である。画面402は、キーワード追加のために用意された画面で、これを用いて問い合わせ概念に新たにキーワードを追加することができる。ボタン403はキーワード追加の決定のボタンであり、チェックボタン404は、問い合わせ概念への追加キーワードを指定するボタンである。リスト405のキーワードが、予測したキーワードであり、リスト406がその重みである。ここで、提示するキーワードは文献を解析することによって予測したもので、検索結果の漏れを少なくするためのキーワードである。これと同様に、検索結果を絞り込むことに適したキーワードを提示する方法もある。そのような絞り込みのための問い合わせ拡張手法の流れを図6に示す。

【0034】

次に、図5により検索結果の表示について説明する。画面501は通常の実験結果の表示画面であり、画面505は、より詳細な情報を含む検索結果の表示画面である。画面501の "Detail Mode" ボタンをマウスでクリックすると、検索結果の詳細画面505に移る。

【0035】

画面501では、リスト502の順位、リスト503の文書ID、リスト504のタイトルを用いて検索結果を表示している。画面505では、横軸507の文書ID及び横軸508の

スコアにより、横軸方向へ検索結果のスコアの高い順に各文書を取り、縦軸506のキーワードにより、各キーワードが検索にどれだけ影響していたかの詳細を確認することができる。要素509は、横軸507の文書IDが示す文書が縦軸506のキーワードの指すものにどの程度影響を受けているかのスコアが表示されている。

【 0 0 3 6 】

図6は、絞り込みのための問い合わせ拡張手法の流れを示す図である。この手法は、従来の問い合わせ拡張とは異なる。それは、従来は問い合わせ概念の脆弱さを補い、検索結果の漏れを少なくすることを目標として問い合わせに追加するキーワードを選出していたが、この手法では、検索結果が膨大であることを受け、それを削減していき目的とする文献を見つけやすくするために、検索結果を絞り込むことを目標として問い合わせに追加すべきキーワードを選出する。この手法では、問い合わせ601と検索対象の文書集合602に対して索引付け603を行い、問い合わせ概念であるクエリーベクトルという内部表現604、及び検索対象の内部表現605を得る。これと同時に、検索対象の文書集合602の文書ごとに、その文書内での単語の共起情報を算出する。この個別に算出した共起情報は個別共起情報606と呼ぶ。以上の処理の後、検索607としてベクトル空間モデルに従いベクトルの比較を行う。その結果が、検索結果の文書集合608である。クエリーベクトルである内部表現604及び検索結果の文書集合608から、共起される単語を個別共起情報606の中から抽出し、それをもとに絞り込むのに適した文書の予測609をする。その結果が、問い合わせ拡張の候補610である。この手法は、検索結果を受けて抽出したものを使うことで、確実に絞り込める単語を抽出することが可能になっている。

【 0 0 3 7 】

次に、図7により検索結果の詳細表示について説明する。画面701は、検索結果の表示画面であり、リスト702の順位、リスト703の文書ID、リスト704のタイトルについては、前述の通りである。この画面で、文書IDをマウスでクリックして選択することでその文書に関する詳細を見ることができる。画面705及び画面706がそれである。画面705は、システムがローカルに保持している情報を表示したもので、検索の際に用いたキーワードについては強調表示（図には枠で囲んで

表示)をしたものである。また、画面706は、システムに登録済みのオンライン上の文献データベースを直接参照したもので、表示の際に上記と同様にキーワードの強調を付加したものである。

【0038】

次に、図8により問い合わせの再計算について説明する。画面801は、検索結果の表示画面であり、リスト802の順位、リスト803の文書ID、リスト804のタイトルについては、前述の通りである。チェックボタン805は、その検索結果を新しく問い合わせ概念に追加するか否かの指定用のものである。このチェックボタン805で追加する文書を選択し、マウスで"Recalculate"ボタンをクリックすることにより、問い合わせ概念(問い合わせ用のクエリーベクトル)を再度構築し直すことができる。その結果が、画面806である。画面806の表示は前述の問い合わせ概念の表示と同様のものである。したがって、リスト807のキーワード、リスト808のタグ、リスト809の重みについても前述の通りである。

【0039】

次に、図9によりシステム構成と動作について説明する。システムの構成は、サーバ901上に、検索エンジン、クエリーベクトル編集エンジン及びオンライン辞書を配置し、クライアント902上にはブラウザを配置する。ユーザは、クライアント902上でブラウザを用いることでインターネットを介してサーバ901とのインタラクションを持つ。また、サーバ901は必要に応じて、予めシステムに登録済みのオンライン上のデータベース903にインターネットを介してアクセスする。サーバ901の機能は、CD-ROM、DVD-ROM、MO等の記録媒体に記録したプログラムを読み込むことによって、あるいはネットワークを介してプログラムを読み込むことによって実現できる。

【0040】

動作は、クライアント側で問い合わせ用の情報入力904として、キーワードやテキストなどの問い合わせ用の情報源を入力すると、サーバ901側では、問い合わせ概念の構築905としてクエリーベクトルを生成し、クライアント側へ表示画面を送る。クライアント側では、これを受けてクエリーベクトルの詳細を確認する。その際、キーワードから公共DBへ検索906として、登録してあるデータベー

スに対してキーワード検索を行う。これはサーバを介してオンライン上のデータベースにアクセスすることで行われる。オンライン上のデータベースからの結果を受けて、サーバ側はその詳細情報をクライアントに表示する。

【0041】

クライアント側では、さらに、問い合わせ概念の編集907として、キーワードのタグや重みの変更をする。サーバ側では、修正した問い合わせを再構築908という形で、クエリーベクトルの再計算を行う。クライアント側で、検索909を行うと、サーバ側からは、検索結果の表示910として結果の表示画面が来る。これを受けて、クライアント側では、登録済みのデータベースへの追加情報の検索をかけ、関連情報の表示911として、関連情報の表示画面を得る。また、検索結果の表示910から、検索結果の問い合わせ概念へのフィードバック912として、検索結果の中から問い合わせ概念に追加する文書を選択することができる。これを受けて、最後にユーザによる再検索913が行われることで、フィードバックも実現する。再検索913以降は、基本的に検索909以降と同様である。

【0042】

【発明の効果】

本発明によれば、データベースからの文献検索において様々な要求を問い合わせとして指定することができ、同時に検索結果の文書からのフィードバックも様々な手法で行うことができる。また、検索結果からさらに、登録済みのデータベースへの検索を行うことが可能になる。

【図面の簡単な説明】

【図1】

検索システムの初期画面である問い合わせ作成のメイン画面を示す図。

【図2】

問い合わせ概念の表示画面例を示す図。

【図3】

問い合わせ概念の詳細を確認する流れを示す図。

【図4】

問い合わせ概念へのキーワードの追加の様子を示す図。

【図 5】

検索結果、及びその詳細を示す図。

【図 6】

絞り込みのための問い合わせ拡張の流れを示す図。

【図 7】

検索結果の文献内容表示画面を示す図。

【図 8】

問い合わせの再計算への流れを示す図。

【図 9】

システム構成と動作を示す図。

【符号の説明】

101…問い合わせ概念の生成用画面

108…問い合わせ概念の表示画面

201…問い合わせ概念の表示例

208…問い合わせ概念の表示例

402…キーワード追加画面

501…検索結果の表示画面例

502…順位のリスト

503…文書IDのリスト

504…タイトルのリスト。

505…検索結果の詳細表示例

701…検索結果の表示画面

705…システムがローカルに保持している文献内容を表す画面

706…オンライン上の文献データベースを直接参照した文献内容を表す画面

901…サーバ

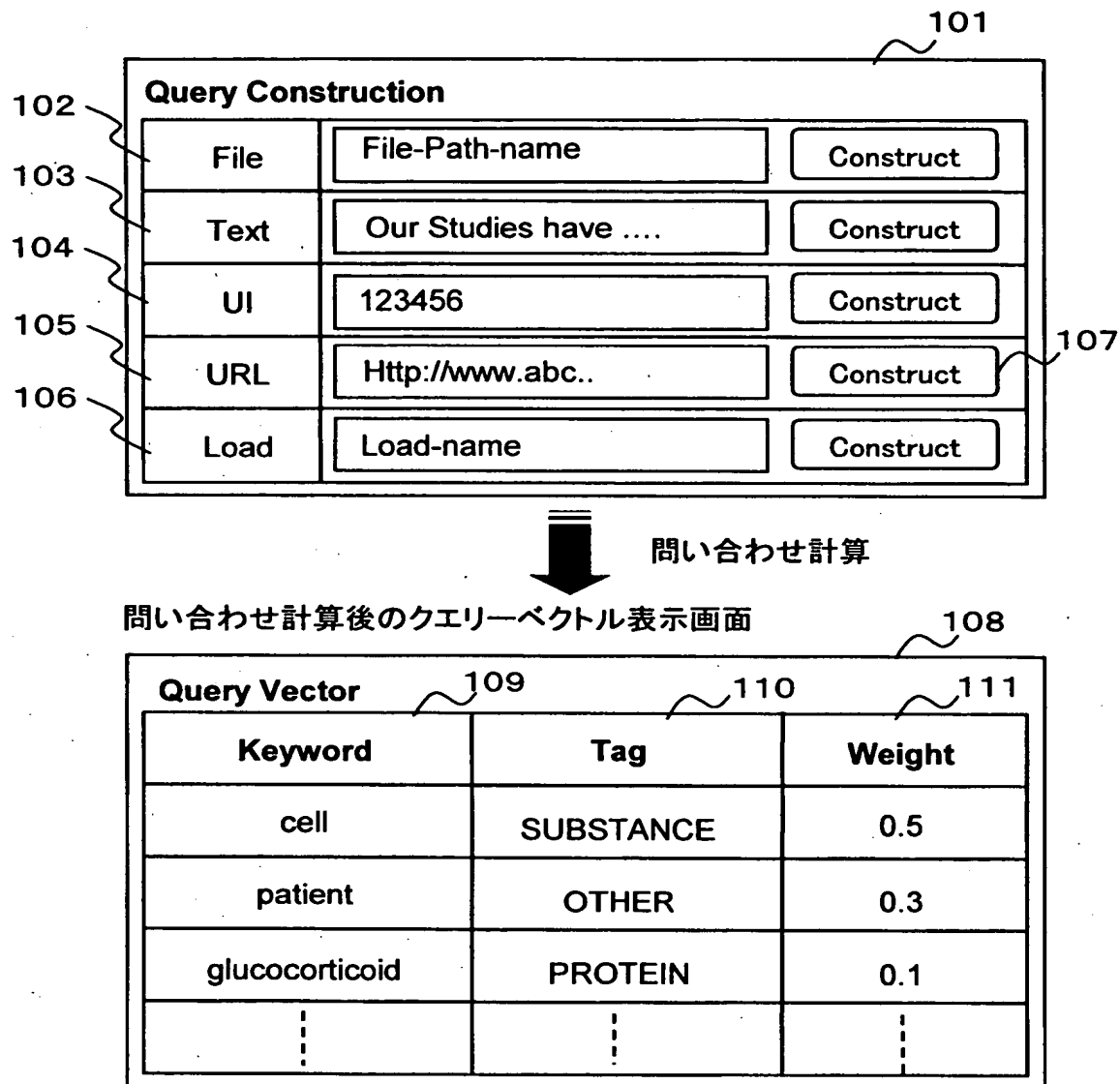
902…クライアント

903…オンライン上のデータベース

【書類名】 図面

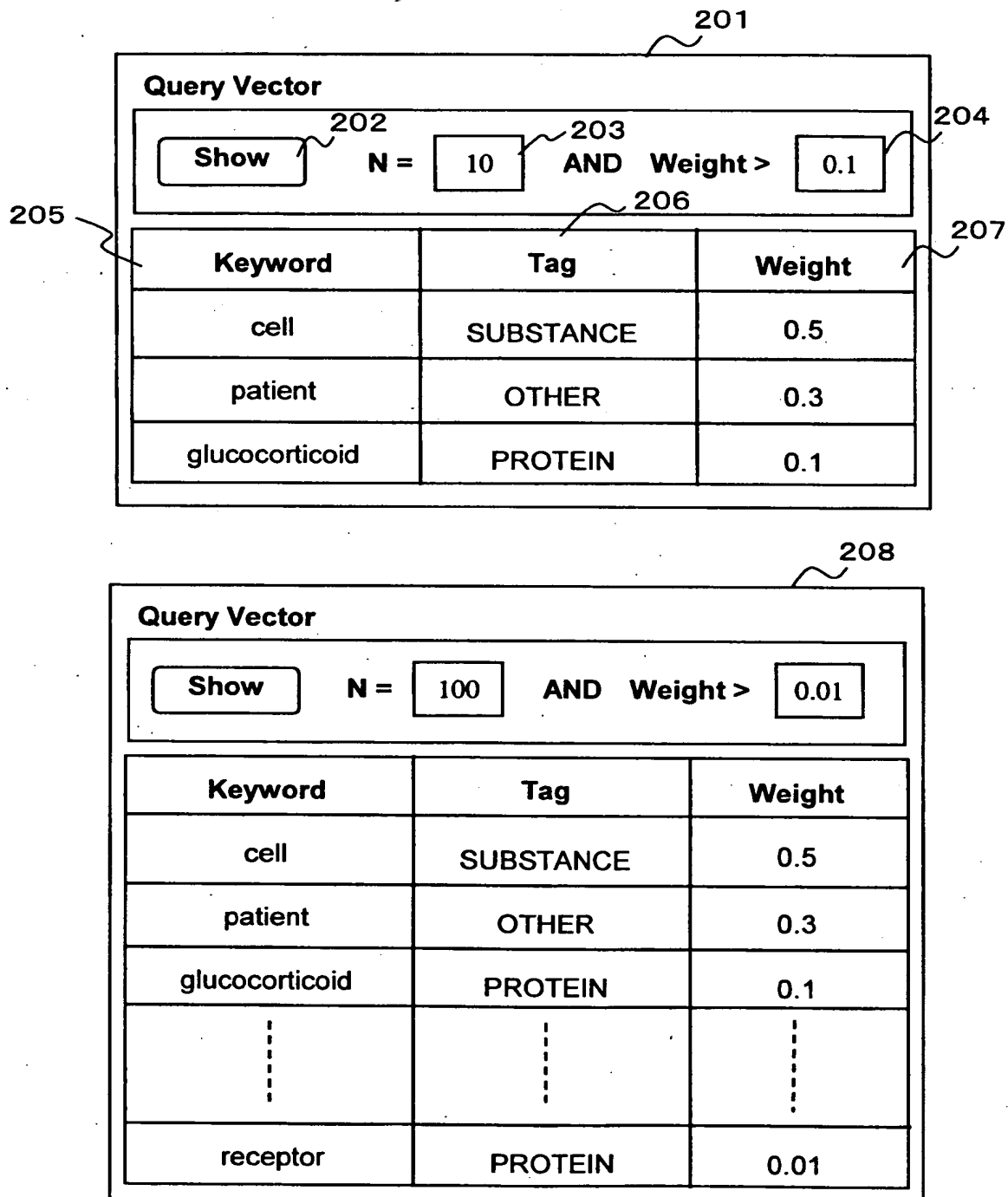
【図 1】

図 1



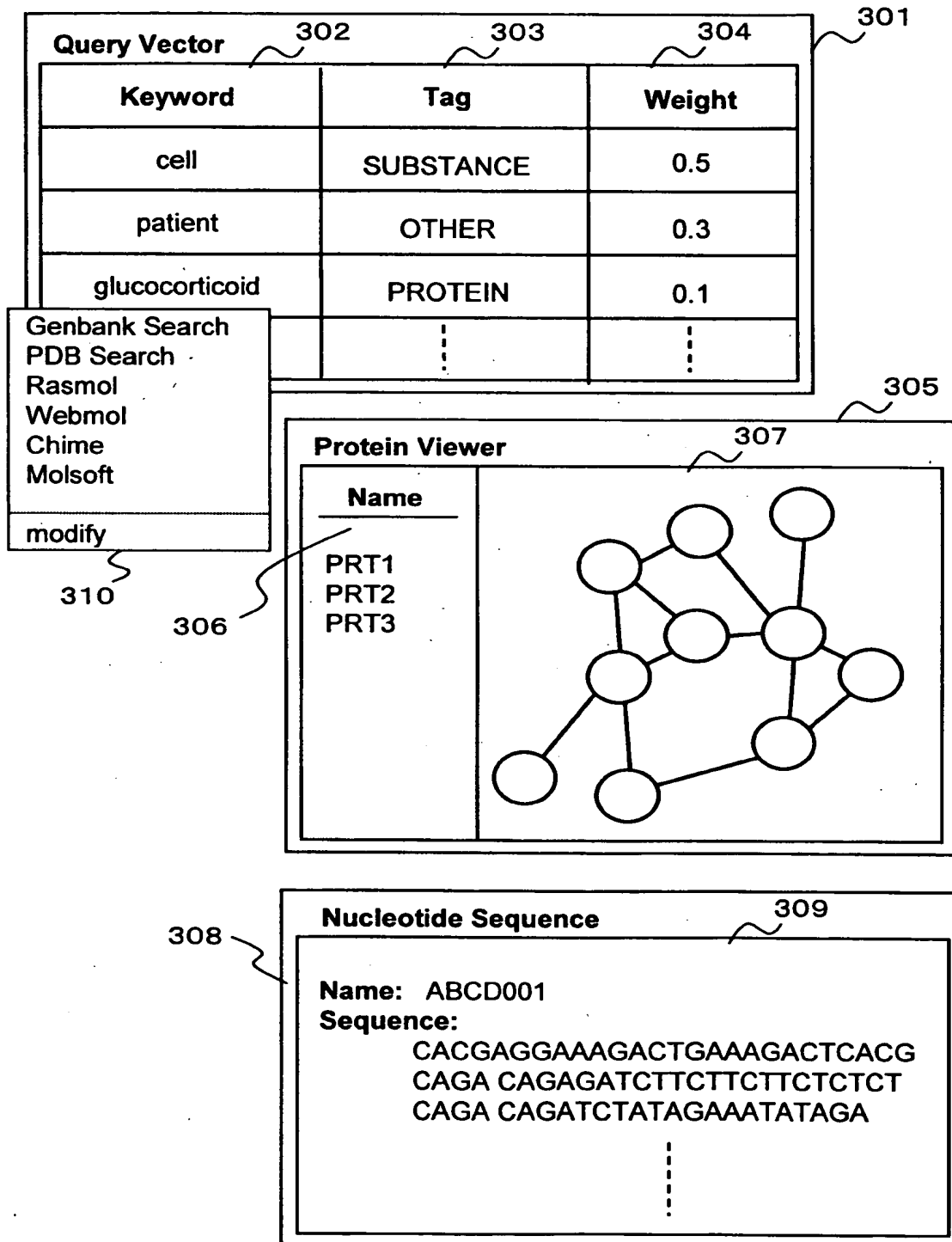
【図 2】

図2



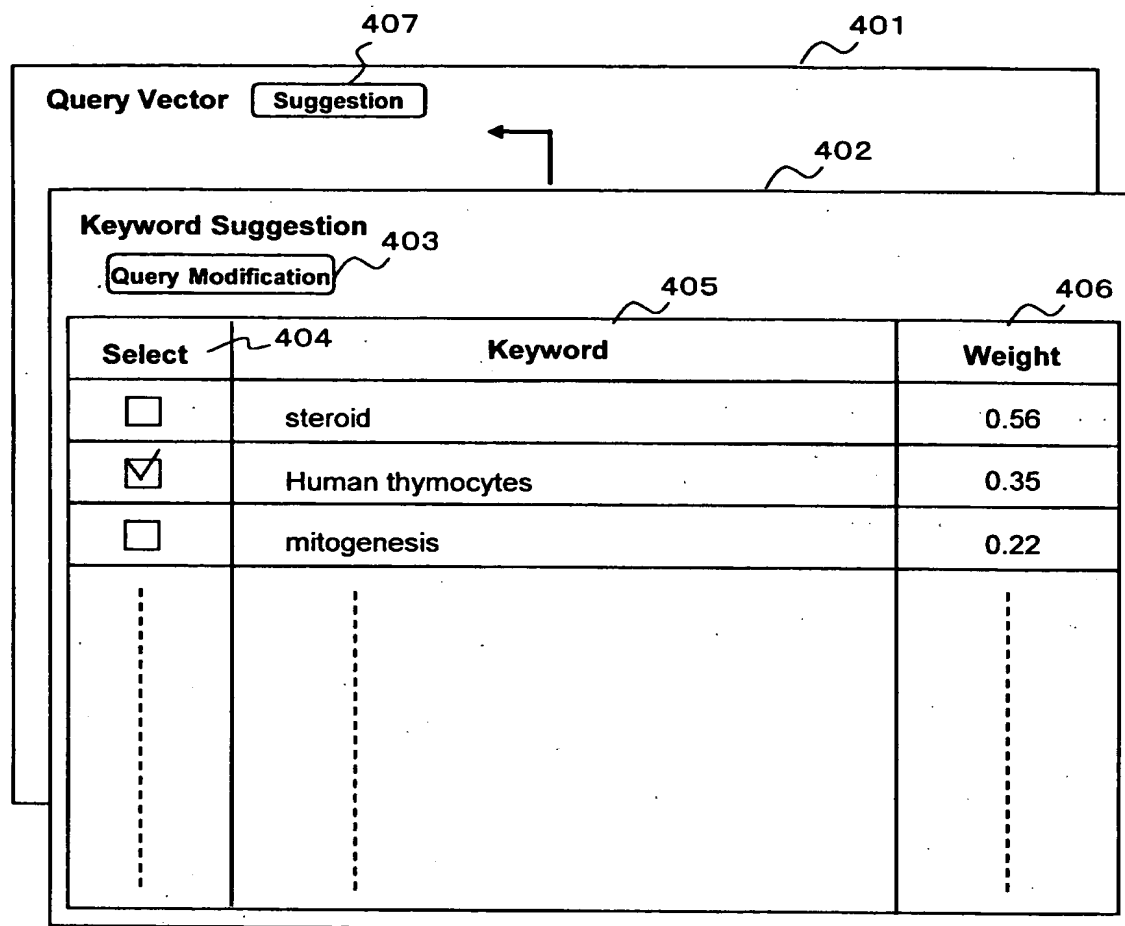
【図 3】

図3



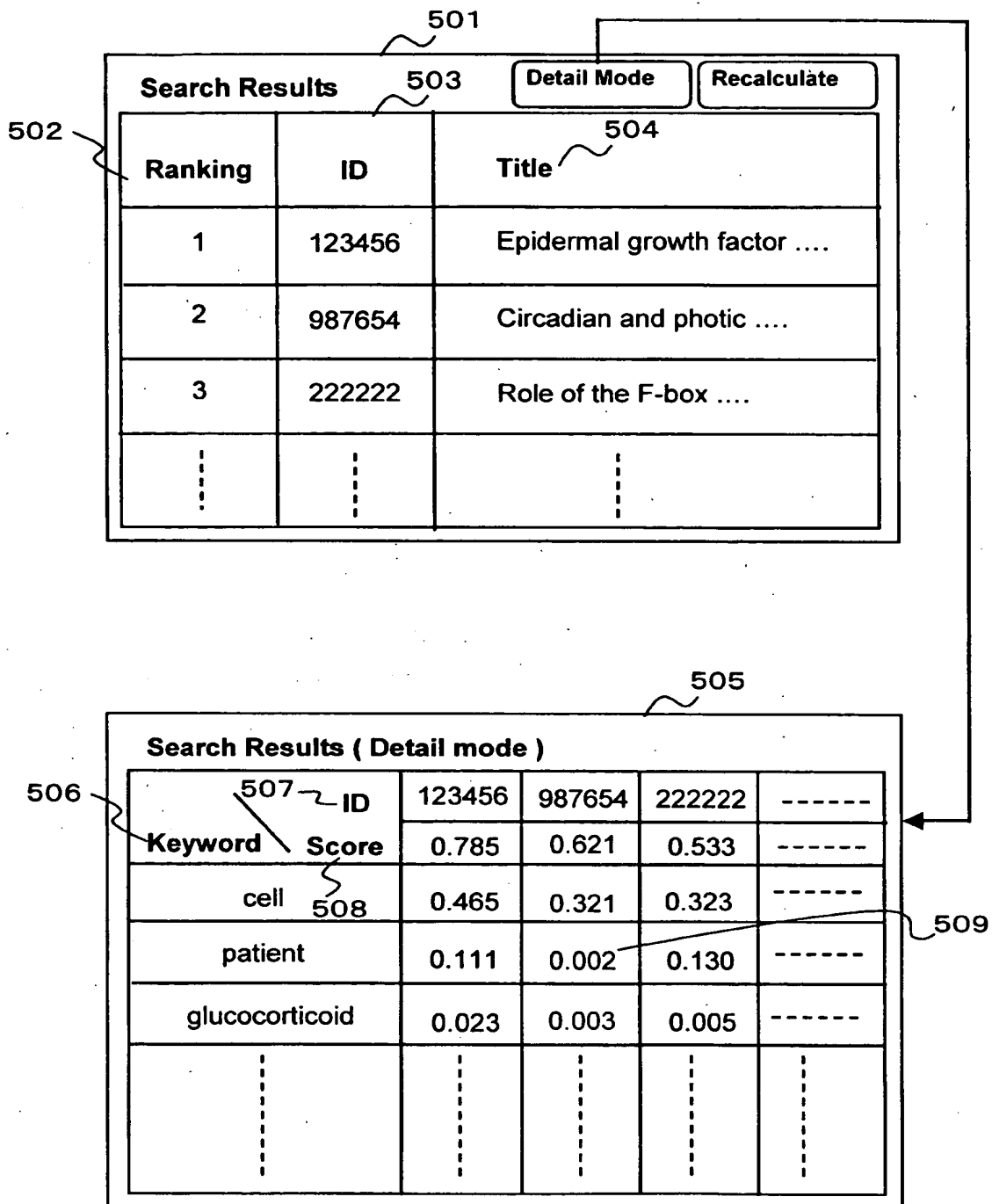
【図 4】

図 4



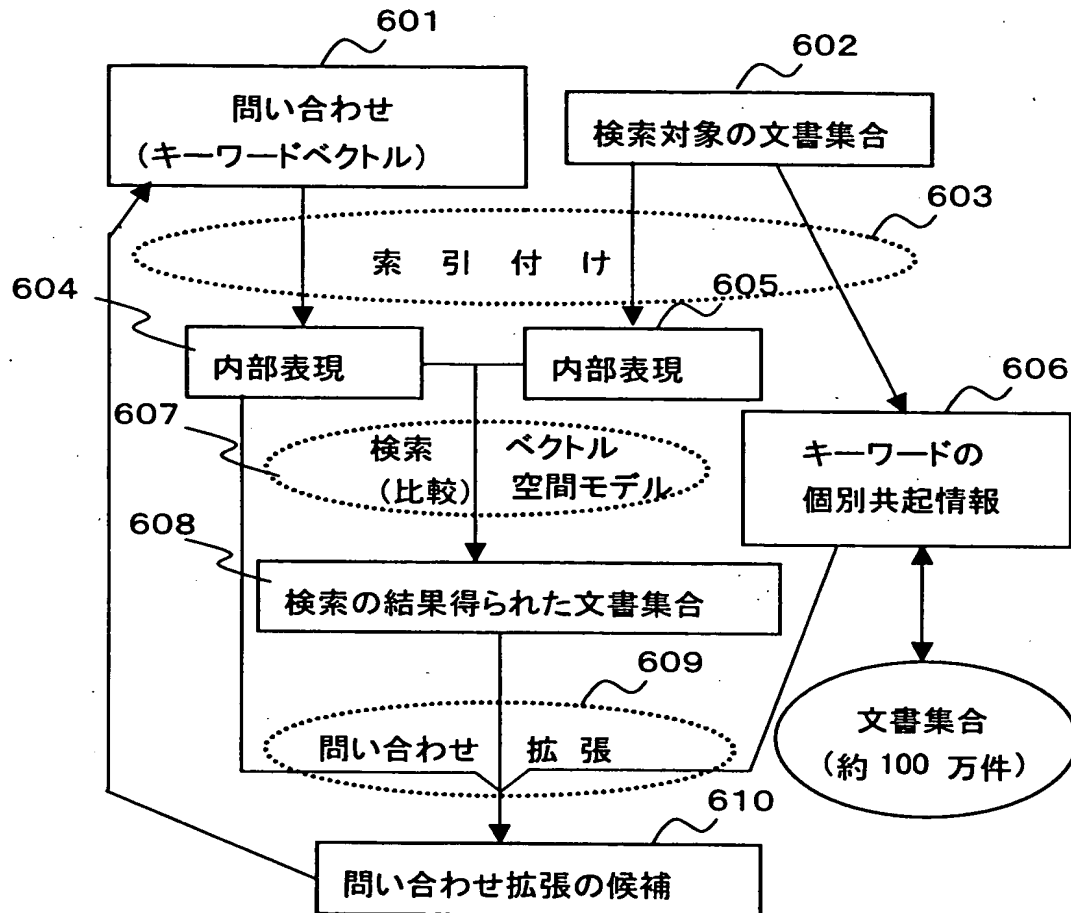
【図 5】

図5



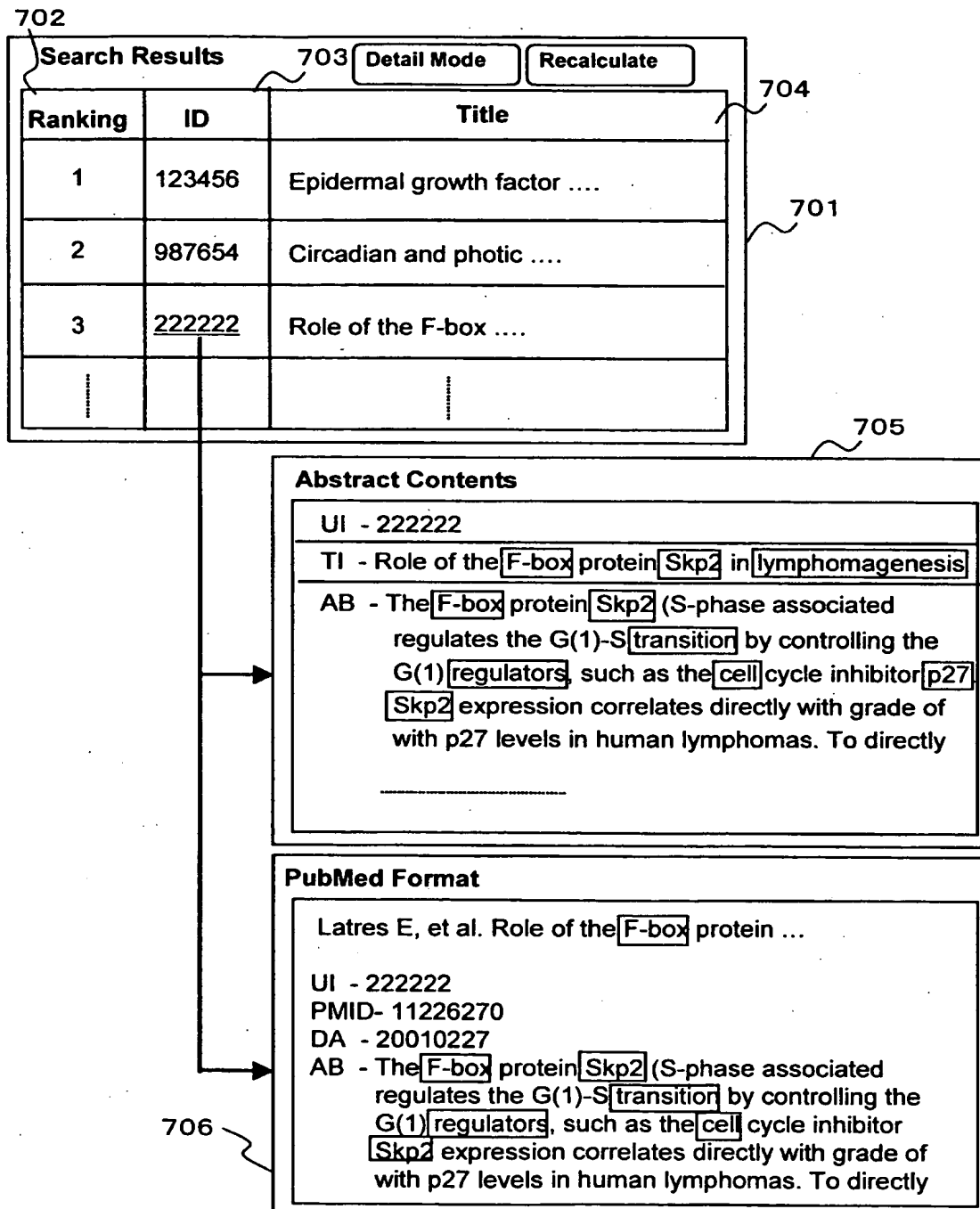
【図 6】

図6



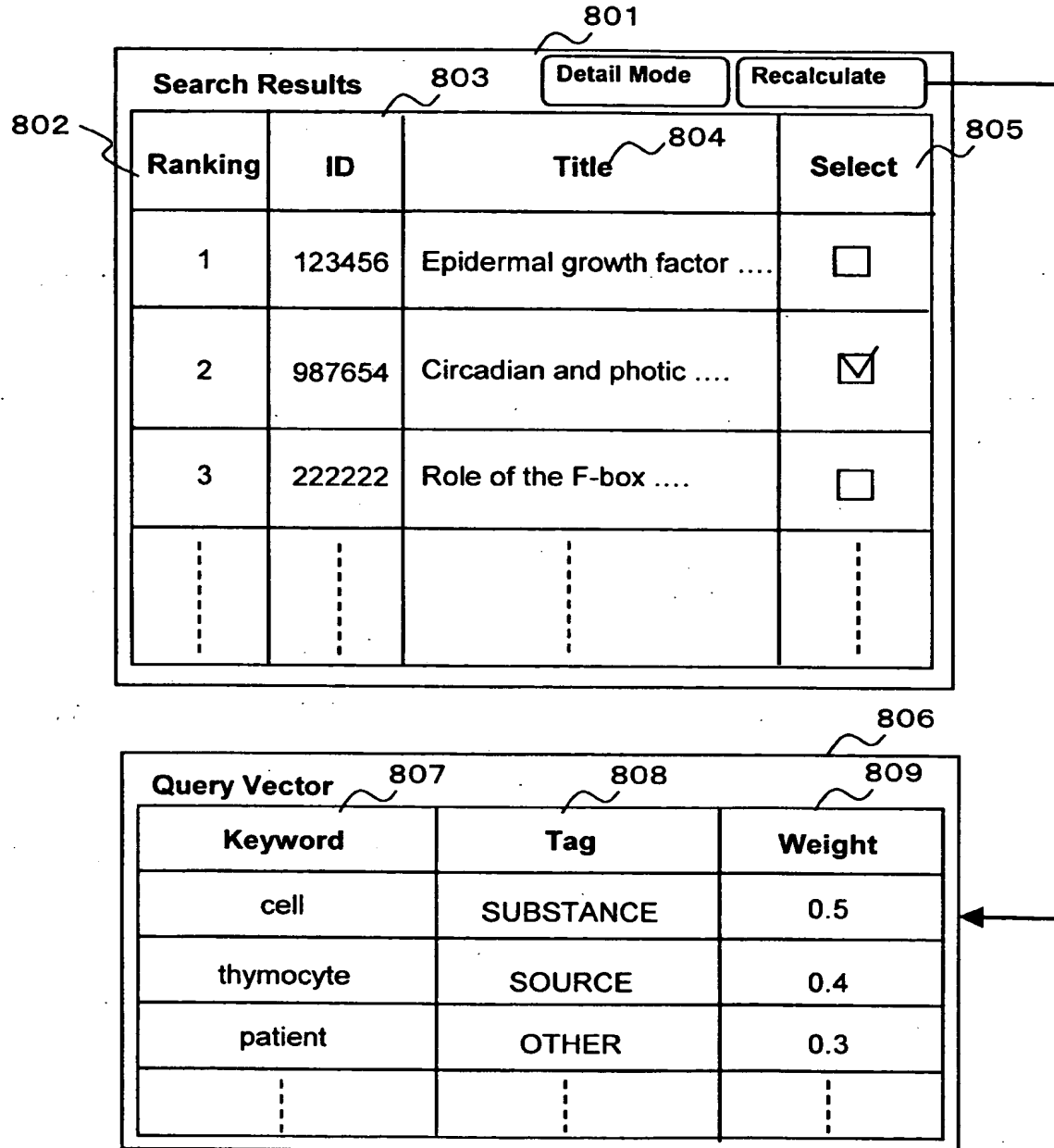
【図 7】

図 7

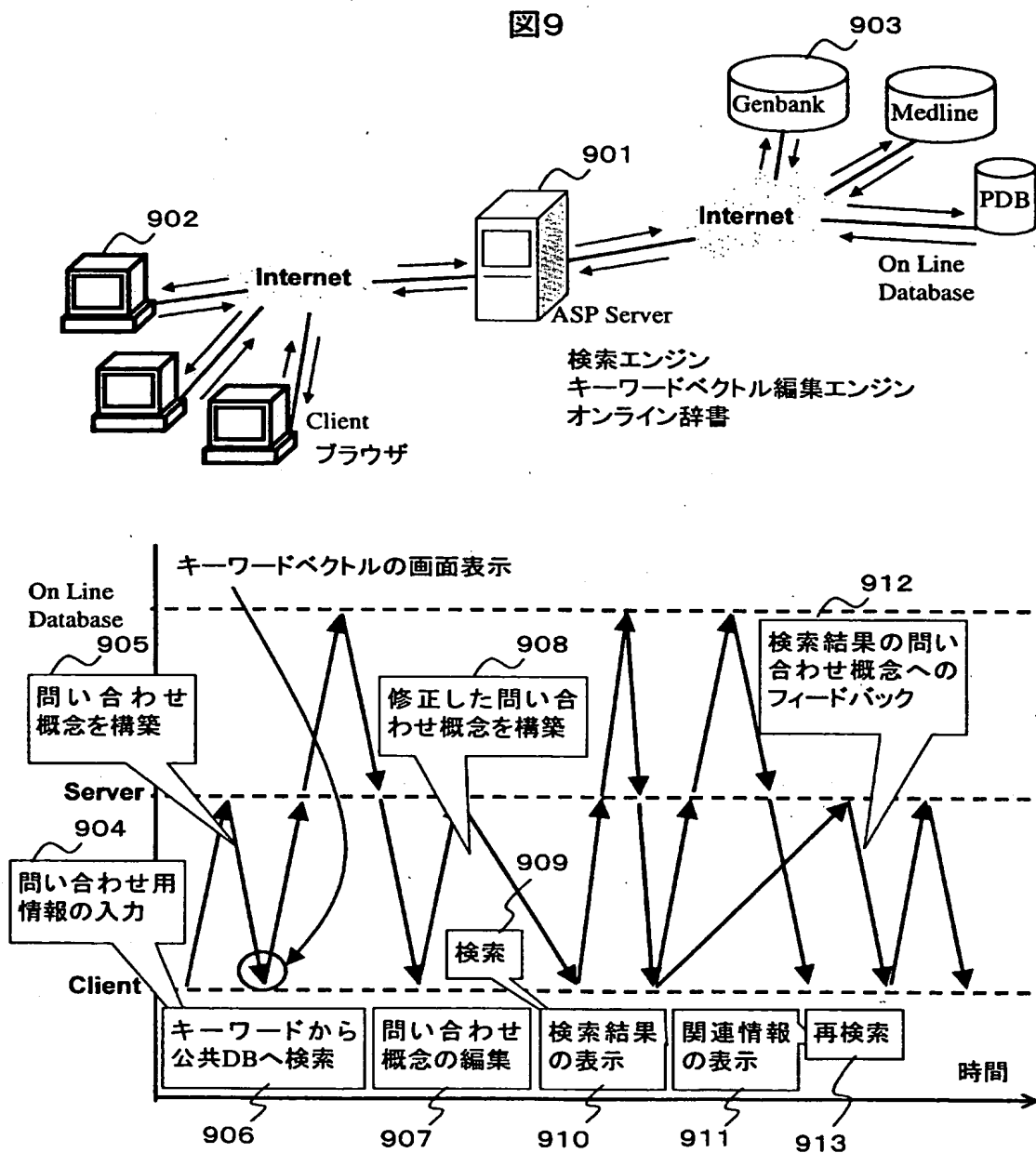


【図 8】

図 8



【図 9】



【書類名】 要約書

【要約】

【課題】 ユーザの要求する情報をよりの確に、より分かりやすく提供する。

【解決手段】 問い合わせ概念生成用画面 1 0 1 でテキスト形式ファイル名入力用フォーム 1 0 2、自然言語入力用フォーム 1 0 3、UI 番号入力用フォーム 1 0 4、URL 入力用フォーム 1 0 5、登録済みの問い合わせの概念読み出し用フォーム 1 0 6 等によって問い合わせ用の情報を入力すると、問い合わせ用の情報から構築した問い合わせ概念を複数のキーワードと各キーワードの重みとを含むクエリーベクトルとして画面 1 0 8 に表示する。ユーザはクエリーベクトルを見て問い合わせ概念を確認し、必要があれば修正することができる。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000005108]

1. 変更年月日	1990年 8月31日
[変更理由]	新規登録
住 所	東京都千代田区神田駿河台4丁目6番地
氏 名	株式会社日立製作所